Lecture 15: Comparing Microarray Analysis Methods

- Comparing ALL and MLL
- The ClassComparison Package
- Comparing processing methods
  - Shedden et al.
  - Wilcoxon rank-sum tests
  - Thresholds
  - Cope et al.
Comparing ALL and MLL

Early in the course (and in the latest homework), we looked at the ALL-MLL-AML data from the paper by Armstrong et al. in *Nature Genetics*, 2002; 30:41-47.

We learned that the AML samples were run on the U95Av2 chip, while the ALL and MLL samples were run on the U95A. We processed all the data in dChip (using a probe mask file to combine the two array types).

To analyze the data, we clustered the samples. We also found 628 differentially expressed genes between ALL and MLL.

We are now going to export the data from dChip and load it into R.
Exporting all the data from dChip

1. Use menu “Tools” – > “Export Expression Value”.

2. Select “all genes”

3. Press “OK”
Starting in R

Load the \texttt{affy} package.

\begin{verbatim}
> require(affy)
Loading required package: affy
Loading required package: Biobase
Loading required package: tools
Welcome to Bioconductor
  Vignettes contain introductory material. To view, simply type: openVignette()
  For details on reading vignettes, see the openVignette help page.
Loading required package: reposTools
[1] TRUE
\end{verbatim}
Load the Sample Information file

> # remember where the data lives
> home <- 'g:/ShortCourse'
> # use the same sample info file we made for dChip
> si <- read.table(file.path(home, 'InfoFiles',
+   'krc-sample-info.xls'),
+   header=TRUE, sep='\t')
> si[1:5,]

<table>
<thead>
<tr>
<th>Scan.name</th>
<th>type</th>
<th>Split</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL01</td>
<td>CL20010111101AA</td>
<td>ALL</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Training</td>
</tr>
<tr>
<td>ALL02</td>
<td>CL20010111104AA</td>
<td>ALL</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Training</td>
</tr>
<tr>
<td>ALL03</td>
<td>CL20010111105AA</td>
<td>ALL</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Training</td>
</tr>
<tr>
<td>ALL04</td>
<td>CL20010111108AA</td>
<td>ALL</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Training</td>
</tr>
<tr>
<td>ALL05</td>
<td>CL20010111109AA</td>
<td>ALL</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Training</td>
</tr>
</tbody>
</table>
Load dChip’s array file

> arrays <- read.table(file.path(home, 'Output',
> 'affyShortCourse arrays.xls'),
> + header=TRUE, as.is=TRUE, sep='\t')
> # Fix the column names!
> dimnames(arrays)[[2]] <-
> + c(dimnames(arrays)[[2]][2:8], 'x')
> # Use the sample name as the row name
> dimnames(arrays)[[1]] <- arrays$Array
> # Only keep useful columns
> arrays <- arrays[, 3:6]
> # Give them sensible names
> dimnames(arrays)[[2]] <- c('MedianIntensity',
> + 'PercentPresent', 'ArrayOutlier',
> + 'SingleOutlier')
Combine sample information

> # Merge sample info with dChip info
> si <- merge(si, arrays, by='row.names',
+    sort=FALSE)
> # Sigh. Fix the row names yet again.
> dimnames(si)[[1]] <- si$Row.names
> # Remove redundant columns
> si <- si[, 2:8]
> rm(arrays) # cleanup
Note that the order has changed!

```r
> si[1:5,]

<table>
<thead>
<tr>
<th>Scan.name</th>
<th>type</th>
<th>Split</th>
<th>MedianIntensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL01</td>
<td>CL20010111101AA</td>
<td>ALL</td>
<td>Training</td>
</tr>
<tr>
<td>ALL24</td>
<td>CL20010111102AA</td>
<td>ALL</td>
<td>Test</td>
</tr>
<tr>
<td>ALL02</td>
<td>CL20010111104AA</td>
<td>ALL</td>
<td>Training</td>
</tr>
<tr>
<td>ALL03</td>
<td>CL20010111105AA</td>
<td>ALL</td>
<td>Training</td>
</tr>
<tr>
<td>ALL04</td>
<td>CL20010111108AA</td>
<td>ALL</td>
<td>Training</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PercentPresent</th>
<th>ArrayOutlier</th>
<th>SingleOutlier</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL01</td>
<td>48.2</td>
<td>1.648</td>
</tr>
<tr>
<td>ALL24</td>
<td>38.3</td>
<td>3.778</td>
</tr>
<tr>
<td>ALL02</td>
<td>49.5</td>
<td>1.450</td>
</tr>
<tr>
<td>ALL03</td>
<td>36.8</td>
<td>3.152</td>
</tr>
<tr>
<td>ALL04</td>
<td>38.7</td>
<td>5.299</td>
</tr>
</tbody>
</table>
```
Specialized factors

> # make a factor to compare ALL vs MLL
> temp <- si$etype
> temp[temp=='AML'] <- NA
> si$ALLvMLL <- factor(temp)
>
> temp <- si$etype
> temp[temp=='MLL'] <- NA
> si$ALLvAML <- factor(temp)
>
> temp <- si$etype
> temp[temp=='ALL'] <- NA
> si$MLLvAML <- factor(temp)
> temp <- si$type
> temp[temp=='AML'] <- 'Other'
> temp[temp=='MLL'] <- 'Other'
> si$ALLvOther <- factor(temp)
>
> temp <- si$type
> temp[temp=='ALL'] <- 'Other'
> temp[temp=='MLL'] <- 'Other'
> si$AMLvOther <- factor(temp)
>
> temp <- si$type
> temp[temp=='AML'] <- 'Other'
> temp[temp=='ALL'] <- 'Other'
> si$MLLvOther <- factor(temp)
>
> si$type <- factor(si$type)
> summary(si)

<table>
<thead>
<tr>
<th>Scan.name</th>
<th>type</th>
<th>Split</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length:72</td>
<td>ALL:24</td>
<td>Length:72</td>
</tr>
<tr>
<td>Class :character</td>
<td>AML:28</td>
<td>Class :character</td>
</tr>
<tr>
<td>Mode :character</td>
<td>MLL:20</td>
<td>Mode :character</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MedianIntensity</th>
<th>PercentPresent</th>
<th>ArrayOutlier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.  : 804</td>
<td>Min.  :28.30</td>
<td>Min.  : 0.253</td>
</tr>
<tr>
<td>1st Qu.:1222</td>
<td>1st Qu.:36.80</td>
<td>1st Qu.: 0.729</td>
</tr>
<tr>
<td>Median :1442</td>
<td>Median :41.10</td>
<td>Median : 1.085</td>
</tr>
<tr>
<td>Mean  :1483</td>
<td>Mean  :40.46</td>
<td>Mean  : 1.724</td>
</tr>
<tr>
<td>3rd Qu.:1727</td>
<td>3rd Qu.:44.83</td>
<td>3rd Qu.: 1.697</td>
</tr>
<tr>
<td>Max.  :3097</td>
<td>Max.  :49.80</td>
<td>Max.  :14.337</td>
</tr>
<tr>
<td>Test</td>
<td>ALLvMLL</td>
<td>ALLvAML</td>
</tr>
<tr>
<td>----------------------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>Min.</td>
<td>0.0440</td>
<td>24</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>0.1610</td>
<td>20</td>
</tr>
<tr>
<td>Median</td>
<td>0.2405</td>
<td>NA’s:28</td>
</tr>
<tr>
<td>Mean</td>
<td>0.2741</td>
<td>NA’s:24</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>0.3460</td>
<td></td>
</tr>
<tr>
<td>Max.</td>
<td>0.9520</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test</th>
<th>ALLvOther</th>
<th>AMLvOther</th>
<th>MLLvOther</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>24</td>
<td>28</td>
<td>20</td>
</tr>
<tr>
<td>AML</td>
<td>48</td>
<td>44</td>
<td>52</td>
</tr>
<tr>
<td>MLL</td>
<td>48</td>
<td>44</td>
<td>52</td>
</tr>
</tbody>
</table>
Create the phenoData object

```r
> pd <- new('phenoData', pData=si, varLabels=list(
+   Scan.name='CEL file name',
+   type='Histological classification',
+   Split='Used as training or test,
+   MedianIntensity='Unnormalized median brightness,
+   PercentPresent='Percentage of present calls',
+   ArrayOutlier='Percentage of Array Outliers',
+   SingleOutlier='Percentage of Single Outliers',
+   ALLvMLL='binary classifier',
+   ALLvAML='binary classifier',
+   MLLvAML='binary classifier',
+   ALLvOtherL='binary classifier',
+   AMLvOther='binary classifier',
+   MLLvOther='binary classifier'))
```
Create the MIAME object

MIAME = minimum information about a microarray experiment

Some of the BioConductor routines require a MIAME object, even though they will let you submit a character string as a description.

> miame <- new('MIAME',
+ name='SA Armstrong',
+ lab='Lander-Golub',
+ title='MLL translocations')
Read in the data from dChip

> temp <- read.table(file.path(home, 'Output',
> 'affyShortCourse expression.xls'),
> header=TRUE, as.is=TRUE, sep='\t',
> quote='', comment.char='')
> # expression data in the later columns
> data <- as.matrix(temp[, 6:77])
> # gene identifiers in the first five columns
> gi <- temp[, 1:5]
> # Use probe sets as row names
> dimnames(gi)[[1]] <- gi$probe.set
> dimnames(data)[[1]] <- gi$probe.set
Check that the order agrees

We noticed that the order of entries in the sample info file had changed when we merged it with the dChip array information. Just to be on the safe side, we should make sure that the order of the data columns matches the sample info rows.

> sum(dimnames(si)[[1]] != dimnames(data)[[2]])
[1] 0
> sum(dimnames(si)[[1]] == dimnames(data)[[2]])
[1] 72
Turn the dChip data into an exprSet

We can bring the dChip quantifications directly into R and turn them into an exprSet. Note that this avoids the memory problems by not bringing in the individual CEL files and not producing an AffyBatch.

```r
> dchip <- new('exprSet',
+     exprs=data,
+     phenoData=pd,
+     annotation='hgu95av2',
+     description=miame,
+     notes='processed by KRC in dChip')
> rm(temp, data, si) # cleanup
```
Meanwhile, we also loaded the CEL files and processed the data using just.rma. Since you will have done that in the homework, I am not going to put the code to do that here. In our analysis, the exprSet created by just.rma was named rmaData.

We are now going to start comparing the results of these two processing methods with respect to finding differentially expressed genes.
The ClassComparison Package

> require(ClassComparison)
Loading required package: ClassComparison
Loading required package: splines
Loading required package: oompaBase
Loading required package: PreProcess
Creating a new generic function for 'plot' in 'PreProcess'
Creating a new generic function for 'print' in 'PreProcess'
Creating a new generic function for 'as.data.frame' in 'PreProcess'
[1] TRUE
T-test, take one

> notAML <- pd@pData$type != 'AML'
> dchip.t <- MultiTtest(dchip[, notAML], 'ALLvMLL')
> dchip.b <- Bum(dchip.t@p.values)
> hist(dchip.b, main='P-values using dChip')
T-test, take two

```r
> rma.t <- MultiTtest(rmaData[, notAML], 'ALLvMLL')
> rma.b <- Bum(rma.t@p.values)
> hist(rma.b, main='P-values using RMA')
```
Do the two methods agree?

```r
> plot(rma.t@t.statistics, dchip.t@t.statistics)
> abline(0,1, col='blue')
```
How do we tell if the methods agree?

We have seen that the BUM plots for t-tests when we used different quantifications methods look similar. We have also seen that the t-statistics roughly agree, in the sense that they more or less follow the identity line. The haze around that line is rather “fat”, however, which suggests that the exact lists of genes we get with the two methods may not be quite the same. Here’s one difference:

```r
> alpha <- 0.05
> countSignificant(dchip.b, alpha=alpha, by='FDR')
[1] 1520
> countSignificant(rma.b, alpha=alpha, by='FDR')
[1] 2353
```
Try a smaller FDR

> alpha <- 0.01
> countSignificant(dchip.b, alpha=alpha, by='FDR')
[1] 681
> countSignificant(rma.b, alpha=alpha, by='FDR')
[1] 992

There certainly appear to be a lot of differentially expressed genes. However, RMA seems to give us more genes than dChip at the same level of the False Discovery Rate. That already tells us something about the processing methods.
How much do the lists overlap?

> # logical vector: what does dChip find?
> dchip.01 <- selectSignificant(dchip.b, 
+     alpha=alpha, by='FDR')
> # logical vector: what does RMA find?
> rma.01 <- selectSignificant(rma.b, 
+     alpha=alpha, by='FDR')
> # Count the overlap
> sum(dchip.01 & rma.01)
[1] 563
> 563/681
[1] 0.8267254

Only 83% of the 681 genes found by dChip are contained in the larger list of genes found by RMA.
Comparing processing methods

So, the answers are “different”. Can we tell which is “better”?

Comparison of seven methods for producing Affymetrix expression scores based on False Discovery Rates in disease profiling data

Kerby Shedden1, Wei Chen2, Rork Kuick3, Debashis Ghosh2, James Macdonald4, Kathleen R Cho5, Thomas J Giordano5, Stephen B Gruber6, Eric R Fearon6, Jeremy MG Taylor2 and Samir Hanash3

Address: 1Department of Statistics, University of Michigan, Ann Arbor, Michigan, USA, 2Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA, 3Department of Pediatrics, University of Michigan, Ann Arbor, Michigan, USA, 4Comprehensive Cancer Center, University of Michigan, Ann Arbor, Michigan, USA, 5Department of Pathology, University of Michigan, Ann Arbor, Michigan, USA and 6Department of Internal Medicine, University of Michigan, Ann Arbor, Michigan, USA

Email: Kerby Shedden - kshedden@umich.edu; Wei Chen - wchen@umich.edu; Rork Kuick - rkuick@umich.edu; Debashis Ghosh - ghosh@umich.edu; James Macdonald - jmacdon@umich.edu; Kathleen R Cho - kathcho@umich.edu; Thomas J Giordano - giordano@umich.edu; Stephen B Gruber - sgruber@umich.edu; Eric R Fearon - ifearon@umich.edu; Jeremy MG Taylor - jmg@umich.edu; Samir Hanash - shanash@umich.edu

* Corresponding author
They looked at 7 processing methods in two different data sets.

- **Methods**
  - dChip
  - GCRMA-EB
  - GCRMA-MLE
  - MAS5
  - PDNN
  - RMA
  - trimmed mean (TM)

- **Data Sets**
  - 47 Colon cancer, U133A (40 MSS vs. 7 MSI)
  - 79 Ovarian cancer, U133A (38 endometroid vs. 41 serous)
Figure 1
Sensitivity results for colon and ovary data. Top row: number of significant probe sets at a range of \( FDR_0 \) values using the t-test statistic. Bottom row: number of significant probe sets at a range of \( FDR_0 \) values using the rank-sum statistic. The left column shows the results for colon data and the right column shows the results for ovary data.
Same idea, ALL-MLL data set

> alpha <- seq(0, 0.1, by=0.01)
> f <- function(a, data) {
+   countSignificant(data, alpha=a, by='FDR')
+ }
>
> dchip.counts <- sapply(alpha, f, dchip.b)
> dchip.counts
  [1]  0  681  936 1139 1356 1520 1703
  [8] 1850 1990 2126 2266
> rma.counts <- sapply(alpha, f, rma.b)
> rma.counts
  [1]  0  992 1379 1725 2078 2353 2623
  [8] 2890 3112 3383 3618
Making the plot

```r
> plot(alpha, rma.counts,
+     xlab='FDR', ylab='Number of probe sets',
+     main='ALL t-statistic', type='b',
+     pch=16, col='blue')
> lines(alpha, dchip.counts, type='b', pch=16)
> legend(0, 3500, c('RMA', 'dChip'), lwd=3,
+        col=c('blue', 'black'))
```
RMA gives more differences in this data set
Wilcoxon rank-sum tests

In the Shedden paper, they use a rank-sum statistic to test for differential expression, in addition to the t-statistic. To compute these statistics, we use the `MultiWilcoxonTest` function in the `ClassComparison` package.

```r
> dchip.wil <- MultiWilcoxonTest(dchip[, notAML],
+   'ALLvMLL')
> rma.wil <- MultiWilcoxonTest(rmaData[, notAML],
+   'ALLvMLL')
```
Summary plots from the Wilcoxon empirical Bayes

> opar <- par(mai=c(0.5, 0.7, 0.2, 0.2),
+     mfrow=c(2,2))
> hist(dchip.wil, main='dChip')
> plot(dchip.wil, prior=0.725, ylim=c(0,1))
> abline(h=0)
> hist(rma.wil, main='RMA')
> plot(rma.wil, prior=0.56, ylim=c(0,1))
> abline(h=0)
> par(opar)
Why is the RMA version skewed?
Counts as a function of posterior probability

```r
> sig <- seq(1.0, 0.9, by=-0.01)
> f2 <- function(s, p, data) {
+   countSignificant(data, prior=p, signif=s)
+ }
> dchip.w.counts <- sapply(sig, f2, p =0.725,
+   data=dchip.wil)
> dchip.w.counts
 [1] 0 394 481 562 648 695 756 824
 [9] 874 908 971
> rma.w.counts <- sapply(sig, f2, p =0.56,
+   data=rma.wil)
> rma.w.counts
 [1] 0 551 707 842 971 1091 1204 1288
 [9] 1384 1479 1598
```
RMA still gives more differences

ALL rank-sum-statistic

Number of probe sets

Posterior Probability

© Copyright 2004–2006, Kevin R. Coombes and Keith A. Baggerly

GS01 0163: ANALYSIS OF MICROARRAY DATA
Figure 3
Calibration results for ovary and colon data. The threshold test statistic required to obtain a given FDR_0 for each method is plotted against the FDR_0 value. Results are shown for the colon data (left column), the ovary data (right column), and for the t-test statistic (top row), and the rank-sum statistic (bottom row).
> alpha <- seq(0.001, 0.1, by=0.01)
> g <- function(a, data) {
+   pval <- cutoffSignificant(data, alpha=a,
+                           by='FDR')
+   qt(1-2*pval, 70)
+ }
> dchip.cut <- sapply(alpha, g, dchip.b)
> rma.cut <- sapply(alpha, g, rma.b)
> plot(alpha, dchip.cut,
+      xlab='FDR', ylab='t-statistic threshold',
+      main='ALL t-statistic', type='b', pch=16,
+      ylim=c(1,4.15))
> lines(alpha, rma.cut, type='b', pch=16,
+       col='blue')
> legend(0.05, 4, c('RMA', 'dChip'), lwd=3,
+        col=c('blue', 'black'))
ALL T-Statistic Thresholds

ALL t-statistic

- Blue line: RMA
- Black line: dChip

FDR

t-statistic threshold

© Copyright 2004–2006, Kevin R. Coombes and Keith A. Baggerly

GS01 0163: ANALYSIS OF MICROARRAY DATA
ALL Rank-Sum Thresholds: Low

ALL rank-sum-statistic

Low rank-sum cutoff

0.92 0.94 0.96 0.98 1.00
Posterior Probability

RMA
dChip
ALL Rank-Sum Thresholds: High
Calibration determines sensitivity

Shedden et al. found the same relation between sensitivity (the number of probe sets called different at a given FDR level) and calibration (the threshold needed to call a probe set different at a given FDR level) in their data sets that we see in our data sets. Namely, methods that provide greater sensitivity do so by lowering the threshold required to call the statistic significant.

In contrast, they found that dChip and TM consistently performed as well or better than other methods on their two data sets. We, of course, found that RMA is “more sensitive”.

They also found (and we agree) that the choice of processing method has a bigger impact on differential expression than the choice of using a paramteric t-statistic compared to a non-parametric rank-sum statistic.
Cope et al., Bioinformatics, 2004; 20:323-331

A benchmark for Affymetrix GeneChip expression measures

Leslie M. Cope\textsuperscript{1}, Rafael A. Irizarry\textsuperscript{2,4}, Harris A. Jaffee\textsuperscript{2}, Zhijin Wu\textsuperscript{2} and Terence P. Speed\textsuperscript{3}

\textsuperscript{1}Department of Mathematical Sciences, Johns Hopkins University, 104 Whitehead Hall, 3400 North Charles Street, Baltimore, MD 21218, USA, \textsuperscript{2}Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe Street, Baltimore, MD 21205, USA and \\
\textsuperscript{3}Department of Statistics, University of California, Berkeley, 367 Evans Hall, Berkeley, CA 94720, USA

Received on May 9, 2003; revised on July 17, 2003; accepted on August 5, 2003

ABSTRACT

Motivation: The defining feature of oligonucleotide expression arrays is the use of several probes to assay each targeted transcript. This is a bonanza for the statistical geneticist, who can create probe set summaries with specific characteristics. There are now several methods available for summarizing probe level data from the popular Affymetrix GeneChips, but it is difficult to identify the best method for a given setting, offering great opportunity to create probe set summaries with specific characteristics. On the other hand, the researcher with data in hand and a particular question in mind is not necessarily able to identify the best method. Using a spike-in study prepared by Affymetrix and a dilution study by Gene Logic as benchmark data, we have developed a graphical tool for the evaluation and comparison of expression measures on the Affymetrix GeneChip platform (Lockhart et al. 1996).
Bench marking methods using “calibration” data

An alternative approach to comparing the results of different processing methods relies on standard sets of spike-in experiments. The performance measures described by Cope et al, are available on a web site:

- Background
- Data and instructions
- Submission form
- Competition results
  - new assessment (of SPIKE-IN HGU95 and HGU133 studies)
  - original assessment (of DILUTION and SPIKE-IN HGU95 studies)
  - entry comparison tool (beta version)
  - study archives

• They use
  • the GeneLogic dilution study that mixed RNA from liver and CNS tissue in different dilutions and proportions
  • the Affymetrix latin-square spike-in study on U95A arrays
  • the Affymetrix latin-square spike-in study on U133A arrays

One should also note that these data sets can be used to evaluate any Affymetrix processing method, by loading the affycomp package.
MAplot of Latin-square data: MAS5
MAplot of Latin-square data: RMA
Standard deviation across replicate dilution arrays

This figure shows the results for four different methods.
Sensitivity to total amount of RNA

For each method, compute the log ratios (fold changes) between lowest (1.25 µg) and highest (20 µg) concentrations in the dilution experiment.
Observed expression vs. nominal concentration in Latin-square

This figure shows the results for four different methods.
Observed vs. Nominal in Dilution

This figure shows the results for four different methods; they fit regressions to intensity as a function of dilution.
ROC curves: general FC

This figure shows the results for four different methods. In each case, they average the ROC curves for different pairwise comparisons in the Latin-square data.
ROC curves: FC=2

This figure shows the results for four different methods
Observed vs. nominal fold change: RMA

This figure shows the results for four different methods.
Observed vs. nominal fold change: RMA, close-up

This figure shows the results for four different methods.
# Measures of performance

<table>
<thead>
<tr>
<th>Assessment</th>
<th>Figure</th>
<th>MAS 5.0</th>
<th>dChip</th>
<th>RMA</th>
<th>Not multi-array</th>
<th>Not robust</th>
<th>MM as PM</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Median SD</td>
<td>2</td>
<td>0.29</td>
<td>0.089</td>
<td>0.088</td>
<td>0.19</td>
<td>0.092</td>
<td>0.074</td>
</tr>
<tr>
<td>(2) R2</td>
<td>2</td>
<td>0.89</td>
<td>0.99</td>
<td>0.99</td>
<td>0.98</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>(3) 1.25v20 corr</td>
<td>3</td>
<td>0.73</td>
<td>0.91</td>
<td>0.94</td>
<td>0.87</td>
<td>0.94</td>
<td>0.93</td>
</tr>
<tr>
<td>(4) 2-fold discrepancy</td>
<td>3</td>
<td>1200</td>
<td>40</td>
<td>21</td>
<td>99</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>(5) 3-fold discrepancy</td>
<td>3</td>
<td>330</td>
<td>8</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(6) Signal detect slope</td>
<td>4a</td>
<td>0.71</td>
<td>0.53</td>
<td>0.63</td>
<td>0.65</td>
<td>0.59</td>
<td>0.55</td>
</tr>
<tr>
<td>(7) Signal detect R2</td>
<td>4a</td>
<td>0.86</td>
<td>0.85</td>
<td>0.8</td>
<td>0.81</td>
<td>0.76</td>
<td>0.72</td>
</tr>
<tr>
<td>(8) Median slope</td>
<td>4b</td>
<td>0.85</td>
<td>0.77</td>
<td>0.87</td>
<td>0.86</td>
<td>0.79</td>
<td>0.76</td>
</tr>
<tr>
<td>(9) AUC (FP &lt; 100)</td>
<td>5a</td>
<td>0.36</td>
<td>0.67</td>
<td>0.82</td>
<td>0.69</td>
<td>0.82</td>
<td>0.81</td>
</tr>
<tr>
<td>(10) AFP, call if fc &gt; 2</td>
<td>5a</td>
<td>3100</td>
<td>37</td>
<td>16</td>
<td>220</td>
<td>19</td>
<td>15</td>
</tr>
<tr>
<td>(11) ATP, call if fc &gt; 2</td>
<td>5a</td>
<td>13</td>
<td>11</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>(12) FC=2, AUC (FP &lt; 100)</td>
<td>5b</td>
<td>0.065</td>
<td>0.17</td>
<td>0.54</td>
<td>0.12</td>
<td>0.52</td>
<td>0.55</td>
</tr>
<tr>
<td>(13) FC=2, AFP, call if fc &gt; 2</td>
<td>5b</td>
<td>1400</td>
<td>12</td>
<td>0.5</td>
<td>18</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>(14) FC=2, ATP, call if fc &gt; 2</td>
<td>5b</td>
<td>3.7</td>
<td>1.3</td>
<td>1.7</td>
<td>2.3</td>
<td>1.4</td>
<td>1.4</td>
</tr>
<tr>
<td>(15) IQR</td>
<td>6</td>
<td>2.7</td>
<td>0.45</td>
<td>0.31</td>
<td>0.67</td>
<td>0.31</td>
<td>0.25</td>
</tr>
<tr>
<td>(16) Obs-intended-loc slope</td>
<td>6a</td>
<td>0.69</td>
<td>0.52</td>
<td>0.61</td>
<td>0.64</td>
<td>0.58</td>
<td>0.54</td>
</tr>
<tr>
<td>(17) Obs-(low)/int-loc slope</td>
<td>6b</td>
<td>0.65</td>
<td>0.32</td>
<td>0.36</td>
<td>0.45</td>
<td>0.34</td>
<td>0.21</td>
</tr>
</tbody>
</table>

The second column denotes the Figure to which the summary statistic relates. Columns 3, 4 and 5 compare MAS 5.0, dChip and RMA. The statistics are described in the text. For each row, the best performing expression measure is denoted with a bold number. Columns 6, 7 and 8 compare RMA to alternatives based on RMA. For each row, if the best performing expression measure is not RMA it is denoted with a bold number.